Adversarial Robustness of Mistral-7B-Instruct-v0.2 and Gemma-7B on University-Level Mathematics

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How robust are Mistral-7B-Instruct-v0.2 and Gemma-7B to adversarial perturbations in problem statements within the MathOdyssey university-level mathematics subset. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Cutting Through the Noise: Boosting LLM Performance on Math Word Problems. Research question: How robust are Mistral-7B-Instruct-v0.2 and Gemma-7B to adversarial perturbations in problem statements within the MathOdyssey university-level mathematics subset?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

16 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 4.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Gemini-1.5 Pro achieves the best overall performance in zero-shot, one-shot, and two-shot inference experiments.	×	0.03
Qwen-1.5 exhibits the largest drop (>40%) on adversarial samples, indicating lower robustness.	×	0.06
Gemini-1.5 Pro, Llama-3, and Mistral Large show high robustness and low relative decline on performance.	×	0.08
Llama-2 (13B) is the best performer in fine-tuning experiments.	×	0.12
Training on adversarial samples boosts average performance on adversarial Simple problems and both Og and adversarial Co	×	0.11
The optimal fine-tuning setting varies across test sets.	×	0.05
The weight of the bags is not relevant to the question in the sample problem about the difference in the number of bags	×	0.04
Acidity level is not a variable that has any relation with the question regarding the number of Lemon heads and boxes in	×	0.01

References

- http://arxiv.org/abs/2412.03205v4
- http://arxiv.org/abs/1807.09380v3
- http://arxiv.org/abs/2406.15444v5